NISTIR XXXX

Developing and Using a System of Risk Tiers for the NIST AI Risk Management Framework

Jeanna Matthews
Patrick Hall
Lori Perine
Reva Schwartz

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.XXXX

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR XXXX

Developing and Using a System of Risk Tiers for the NIST AI Risk Management Framework

Jeanna Matthews
National Institute of Standards and Technology
Information Technology Laboratory
& Clarkson University

Patrick Hall
National Institute of Standards and Technology
Information Technology Laboratory
& The George Washington University

Lori Perine
National Institute of Standards and Technology
Information Technology Laboratory
& The University of Maryland

Reva Schwartz
National Institute of Standards and Technology
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.XXXX

December 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology
Interagency or Internal Report XXXX
Natl. Inst. Stand. Technol. Interag. Intern. Rep. XXXX, 21 pages (December 2023)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.XXXX

Abstract

Organizations using the NIST AI Risk Management Framework (AI RMF) are encouraged to define and develop reasonable risk tolerance for their specific use cases of AI. The AI RMF states "Organizations should follow existing regulations and guidelines for risk criteria, tolerance, and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or established documentation, reporting, and disclosure requirements. Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations should define reasonable risk tolerance. Once tolerance is defined, this AI RMF can be used to manage risks and to document risk management processes." This document discusses risk tolerances and describes a practical process for defining reasonable risk tiers based on the severity of possible impacts and the likelihood of those impacts occurring. It guides organizations to establish a system of risk tiers related to their specific context of use. And it surveys existing risk tier approaches. This document is intended to assist organizations seeking to define and develop reasonable risk tolerance, risk tiers, and related materials and practices.

Key words

AI; Governance; Trustworthiness; Risk tiers; Risk tolerance

Table of Contents

List of Tables

List of Figures

Glossary

control: Measure that is modifying risk [4].

impact: The force of impression of one thing on another; a significant or major effect [5].

level of risk: Magnitude of a risk or combination of risks, expressed in terms of the combination of impacts and their likelihood [4].

likelihood: Chance of something happening [6].

probability: Measure of the chance of occurrence expressed as a number between 0 and 1, where 0 is impossibility and 1 is absolute certainty [4].

risk: Effect of uncertainty on objectives [4]. When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence (Adapted from: OMB Circular A-130:2016). Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet [5].

risk control: Mechanisms at the design, implementation, and evaluation stages [that can be taken] into consideration when developing responsible AI for organizations that includes security risks (cyber intrusion risks, privacy risks, and open source software risk), economic risks (e.g., job displacement risks), and performance risks (e.g., risk of errors and bias and risk of black box, and risk of explainability) [5].

risk management: Coordinated activities to direct and control an organization with regard to risk [4].

risk matrix: Tool for ranking and displaying risks by defining ranges for impact and likelihood [4].

risk tiering: The rating of risk inherent in the use of individual models, which can benefit a firm's resource allocation and overall risk management capabilities [7].

risk tolerance: Risk tolerance refers to the organization's or AI actor's readiness to bear the risk in order to achieve its objectives. [5].

safety: Property of a system such that it does not, under defined conditions, lead to a state in which human life, health, property, or the environment is endangered; [safety involves reducing both the probability of expected harms and the possibility of unexpected harms] [8].

stakeholder: Individual or organization having a right, share, claim, or interest in a system or in its possession of characteristics that meet their needs and expectations. An individual, group, or organization who may affect, be affected by, or perceive itself to be affected by a decision, activity, or outcome of a project [5].

1. Introduction

The NIST Artificial Intelligence Risk Management Framework (AI RMF) is a voluntary resource to assist organizations in managing the risks posed by AI technology to individuals, groups, communities, society, and the environment. The core precept of the Framework is AI system trustworthiness within a culture of responsible AI practice and use [9]. In the context of the AI RMF, risk refers to the composite measure of an event's probability of occurring and the magnitude or degree of the impacts of the corresponding event. When considering the negative impact of a potential event, risk is a function of 1) the negative impact, or magnitude of harm, that would arise if the circumstance or event occurs and 2) the likelihood of occurrence. Negative impact or harm can be experienced by individuals, groups, communities, organizations, society, the environment, and the planet.

While the AI RMF can be used to prioritize risk, it does not prescribe risk tolerance. Risk tolerance refers to an organization's or AI actor's readiness to bear risk in order to achieve its objectives. Organizational risk tolerance is typically espoused and documented by senior leadership. Different organizations may have varied risk tolerances due to their particular organizational priorities and resource considerations [9]. The process presented in this document can help organizations define a set of customized tiers to manage risk in operational settings, in alignment with an organizational risk tolerance. The level of risk that is acceptable to organizations or society are highly contextual, and may be application and use-case specific. A standard application of organizational risk tolerance is to establish criteria for placing AI systems into risk tiers, and to treat risks that fall into higher level tiers with more aggressive risk management practices. The number of risk tiers as well as the requirements placed on each tier can vary.

Organizations using the AI RMF are encouraged to develop and define reasonable risk tolerance for AI and risk tiers for the specific use cases of AI if they do not already have or use them. Determining and documenting organizational risk tolerances and risk tiers is a foundational element of the AI RMF Map function. In the Map function, organizations are encouraged to establish the context to frame risks related to an AI system, the internal expertise necessary to identify and evaluate these contextual factors, and methods for integrating external voices into the process. The organization's risk tolerance and risk tiers further inform the Measure and Manage functions, in particular the definition and selection of measurement and monitoring mechanisms that can support decisions about deployment and/or continued operation of the AI system ("go"/"no-go" decisions). Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems. The costs and benefits associated with AI system trustworthiness can inform risk tolerance development. To leverage cost/benefit measures, organizations need concrete evidence of those benefits including measures related to fitness for purpose, effectiveness, and reliability. System benefit measures should be objective, repeatable/reproducible, empirically validated, and obtained via a well-defined and transparent process.

Organizational risk tolerance should be set deliberately and explicitly by senior leadership (e.g., CxO, board of directors, etc.) and not on an ad-hoc or implicit basis by engineers or data scientists. For example, a joint publication from The Board of Governors of the Federal Reserve System, the Office of the Comptroller of the Currency (OCC), and the Federal Deposit Insurance Corporation (FDIC) states: “In setting the firm’s risk appetite, the board of directors articulates the firm’s tolerance for disruption considering its risk profile and the capabilities of its supporting operational environment” [10]. An organization’s risk tolerance may also be influenced by legal or regulatory requirements, or by policies and norms established by AI system owners, organizations, industries, communities, or policy makers. Even where laws, regulation, and guidelines exist, organizations will often benefit from defining their own risk tolerance.

Risk tolerance is reflected in the the AI RMF Map 1.5 subcategory (“Organizational risk tolerances are determined and documented”) [9]. This sub-category references organizational risk tolerance and guides organizations to devise risk levels based on the possible impacts and estimated likelihood of impacts for AI systems [11]. In some sectors, organizations can follow a body of existing laws, regulations, and/or guidelines for risk, tolerance, tiers, and response that has been established by organizational, domain, discipline, sector, or professional requirements. These existing governance structures can be used to initiate definition of risk tolerance and risk tiers. Risk management is not a one-time activity, but rather a journey. Accordingly, risk tolerances should be expected to change over time as AI systems, policies, and norms evolve with an organization’s level of maturity in adopting and implementing the AI RMF.

> The NIST AI Risk Management Framework “is intended to be flexible and to augment existing risk practices which should align with applicable laws, regulations, and norms. Organizations should follow existing regulations and guidelines for risk criteria, tolerance, and response established by organizational, domain, discipline, sector, or professional requirements. Some sectors or industries may have established definitions of harm or established documentation, reporting, and disclosure requirements. Within sectors, risk management may depend on existing guidelines for specific applications and use case settings. Where established guidelines do not exist, organizations should define reasonable risk tolerance. Once tolerance is defined, this AI RMF can be used to manage risks and to document risk management processes [9].”

The remainder of this document is intended to assist organizations seeking to define and develop reasonable risk tiers and related practices. Specifically, this document provides:

- practical guidance for organizations seeking to establish risk tiers within their specific use cases and operational contexts.

- illustrative examples of how organizations currently use risk tiers for managing risks posed by AI technology and a landscape survey of existing risk tier approaches for managing risk across different contexts.

Section 2 provides a practical process that organizations can use to construct a risk tiering system specific to their own use cases and operational contexts. It defines risk tiers based on the severity of possible impacts and the likelihood of those impacts occurring. The section also provides examples from standards bodies, and suggested actions for how to operationalize a risk tiering system using the NIST AI RMF. Section 3 describes how a system of risk tiers can be used over time within an organization. Section 4 provides a landscape survey of risk tiers and related risk management practices currently in use. Section 4 describes examples from US federal government agencies such as the US Federal Aviation Administration (FAA), the Food and Drug Administration (FDA), international standards, and examples from private industry. These examples illustrate different risk tier styles to provide inspiration for organizations seeking to establish their own risk tier processes.

2. Establishing a System of Risk Tiers

This section presents a practical, five-step process that organizations can use to construct their own risk tiers. The five steps are as follows:

1. establish a range of failure impact levels from high to low.

2. establish a range of failure likelihood levels from high to low.

3. determine the number of risk tiers desired from high to low.

4. map impact and likelihood levels to risk tiers.

5. specify the risk management activities to be performed for each risk tier.

The subsections below cover each of these steps in detail, with specific examples organizations can use for customization. Step 5 requires the most substantial work on the part of organizations using this process to define their risk tiers and involves aligning the activities described in the NIST AI RMF with specific tiers.

2.1 Step 1: Establish a range of failure impact levels from high to low

The output of this section is a table of impact levels from high severity to low severity. Table 1 shows an example from the Institute of Electrical and Electronics Engineers (IEEE) 1012 standard–*Standard for System, Software, and Hardware Verification and Validation*–that is

Table 1. Sample Impact Levels from IEEE 1012 [1] Annex B, Table B.2

Level	Description
Catastrophic	Loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss.
Critical	Major and permanent injury, partial loss of mission, major system damage, or major financial or social loss.
Marginal	Severe injury or illness, degradation of secondary mission, or some financial or social loss.
Neglible	Minor injury or illness, minor impact on system performance, or operator inconvenience.

designed to be broadly applicable to any software or hardware system [1]. Organizations may also modify the number of levels or the descriptions to fit specific use cases.

Section 4 presents risk management examples that may be useful for organizations that want to customize risk severity/impact levels to a specific use case. The examples in Section 4 are specific to the context of the organizations' purview. For example, the FAA risk tiers in Table 2 show the same set of 4 levels (catastrophic, critical, marginal and negligible), but with the qualitative descriptions customized to be specific to the aviation context. While organizations developing AI risk tolerance are likely dealing with vastly different contextual risks, they can adapt their tiers to reference the likely impacts specific to their own domains.

Table 2. Sample Risk Severity/Impact Levels from FAA [2] Chapter 4, "Assessing Risk"

Level	Description
Catastrophic	Results in fatalities and/or total airframe loss.
Critical	Severe injury or major airframe or property damage.
Marginal	Minor injury or minor airframe or property damage.
Negligible	Less than minor injury or damage.

AI risks can impact organizations and enterprises, as well as individuals, communities, society, and the environment. Approaches for identifying and measuring these impacts and their likelihoods within a specific domain and operational context will benefit from the involvement of an interdisciplinary internal team, and the inclusion of individuals and communities who may be impacted by the AI system if deployed. External engagement activities are best performed early in the AI lifecycle. The Map 1 categories of the AI RMF provide detailed guidance for conducting these activities. Impact levels may also be informed by domain expertise, expert judgement, or incident reports from public databases including the AI Incident Database [12], the AIAAIC [13], and the AI Vulnerability Database [14].

Organizations may start by performing a context analysis to identify possible impacts, then refine further for specific use cases within the domain. Informed by impact measures, and interdisciplinary and multi-stakeholder feedback, impacts can be sorted from most to least severe, and then grouped into rough tiers or levels. This step does not require a likelihood estimate for each impact. These activities are also covered in Map subcategory 3.2 of the AI RMF ("[p]otential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness–as connected to organizational risk tolerance–are examined and documented") [9]. Once deployed, organizations should track, quantify, or document costs and other impacts arising from AI systems to improve impact estimates, enhance future risk assessment exercises, and for additional risk management purposes.

Organizations may also consider the level of support for recovery, restitution, repair or other types of recourse in assessing the severity of impacts. For example, negligible impact would generally have very low/zero recovery time or costs, while recovery for catastrophic impacts would be very difficult (if not impossible) for some impacts (e.g., death), infeasible for others (e.g., irrecoverable financial loss) or otherwise not supported. Where recovery is possible, improvements in processes for quick and effective recovery can be an important way to mitigate the severity of possible impacts. Table 3 presents a sample of impact levels with recovery estimates. Estimates for restitution, repair or other types of recourse that may be more directly applicable than recovery can be appropriate in many cases.

Table 3. Sample Impact Level Definitions Including the Level of Support for Recovery

Level	Description
Catastrophic	• Loss of human life, complete mission failure, loss of system security and safety, or extensive financial or social loss. • Full recovery is impossible or unsupported, or the time to full recovery is prohibitively long. • Partial recovery (50% or more) is infeasible.
Critical	• Major and permanent injury, partial loss of mission, major system damage, or major financial or social loss. • Full recovery is impossible or unsupported, or the time to full recovery is prohibitively long. • Partial recovery (50-80% or more) is achievable and supported. Time to partial recovery is predictable and moderate.
Marginal	• Severe injury or illness, degradation of secondary mission, or some financial or social loss. • Full recovery is achievable and supported, but time to full recovery may be unknown or long. • Partial recovery (80% or more) is achievable and supported. Time to partial recovery is predictable and moderate.
Negligible	• Minor injury or illness, minor impact on system performance, or operator inconvenience. • Full recovery is achievable and supported. Time to full recovery is predictable and small.

2.2 Step 2: Establish a range of failure likelihood levels from high to low

The output of this section is a table of likelihood levels from high probability to low probability. Table 4 shows an example from NIST's *Guide for Conducting Risk Assessments* (Special Publication 800-30) where likelihood levels are described with both qualitative terms, and quantitative ranges/semi-quantitative values [3]. Organizations can use this example and modify it for their specific use cases and operational contexts. In the context of the AI RMF, errors, and accidents are not the only causes of harm to consider or factors to be monitored for impacts. AI systems can cause direct and indirect harms even when they are operating as expected.

Table 4. Sample Likelihood Levels from NIST 800-30 [3] Appendix G

Qualitative Values	Semi-Quantitative Values		Description
Very High	96-100	10	Error, accident, or act of nature is almost certain to occur; or occurs more than 100 times a year.
High	80-95	8	Error, accident, or act of nature is highly likely to occur; or occurs between 10-100 times a year.
Moderate	21-79	5	Error, accident, or act of nature is somewhat likely to occur; or occurs between 1-10 times a year.
Low	5-20	2	Error, accident, or act of nature is unlikely to occur; or occurs less than once a year, but more than once every 10 years.
Very Low	0-4	0	Error, accident, or act of nature is highly unlikely to occur; or occurs less than once every 10 years.

Determining likelihood of impact is closely connected to operational context. Before a system is deployed, organizations can estimate the likelihood of impacts based on information about deployed AI systems in similar contexts, domain expertise, expert judgement, and incident reports from public incident databases. Once AI systems are deployed, organizations are encouraged to measure and document actual rates of occurrence and update likelihood estimates accordingly.

2.3 Step 3: Decide on the number of risk tiers desired from high to low

This section provides guidance on defining and mapping the number of risk tiers, where the highest risk tier number corresponds to the highest level of risk. Organizations may also choose to name risk tiers or reverse the labeling of risk tiers, where the lowest risk tier number corresponds to the highest level risk. Organizations can distinguish between different response levels and required risk management resources by specifying a sufficient number of levels. Practical examples include:

- the FDA's use of 3 levels to determine risks related to medical devices: Class 1, 2 and 3 [15].
- the IEEE 1012 specification of 4 levels to determine software and hardware risks: integrity level 1 through 4 [1].

Organizations, especially smaller organizations, or organizations in early stages of their AI risk management practice, may choose to start with 2 or 3 levels for their risk tiers and revisit to assess if the number of levels are sufficient to differentiate risk within their specific use cases. One common and intuitive labeling of 3 risk tiers is to assign *green* for level 1 (risk is low/negligible), *red* for level 3 (risk is high/severe) and *amber* for level 2 (applications or systems where risk is neither low/negligible nor high/severe). Once systems or applications are assigned to risk tiers, a potential indicator that more or less risk tiers may be needed is a lack of balance or an undesirable distribution of resources, requirements, applications, or systems across risk tiers. Ideally, most organizational AI systems or applications are to assigned to lower risk tiers, while a small number are assigned higher risk tiers. Similarly, it is inadvisable that any one risk tier contains so many applications or systems that risk management requirements become unmanageable or infeasible. The next two sections will provide guidance on how to assign meaning–resources, requirements, applications, or systems–to each risk tier level.

2.4 Step 4: Map impact and likelihood levels to risk tiers

The output of this section is a risk assessment matrix that maps impact levels and likelihood levels onto risk tiers. Specifically, this step involves forming a grid or two-dimensional (2-D) matrix of impact levels and likelihood levels. Higher risk tier levels within the grid map to higher impact levels and higher likelihood levels. Table 5 shows a hypothetical example with 4 impact levels, 5 likelihood levels and 3 risk tiers. Table 5 shows impact levels from high to low along one dimension and likelihood levels from high to low along the other dimension. The highest risk tier is assigned in the upper left. A system with a very high likelihood of catastrophic impacts would warrant the highest level of risk management activity. The risk tier in the upper left is where catastrophic impacts intersects with very high likelihood. The lowest risk tier is in the lower right, when the impacts are negligible and the likelihood of those impacts is very low.

Filling in the two extreme corners of the risk assessment matrix is straightforward. More nuance is necessary when filling in the middle sections of the matrix. Risk tiers can be assigned based on the organization's risk tolerance, organizational assessments of expected benefits and costs, expert judgement, domain expertise, multi-stakeholder feedback, and information relating to AI systems in similar deployment contexts. (Map 3 in the AI RMF may provides some helpful guidance [11].) In the Table 5 example, the highest risk tier is assigned whenever there is a risk of catastrophic impact, even when the likelihood of occurrence is very low. This is reflected by assigning tier 3 (the highest risk tier in this example) across the entire top row. The other 3 rows (Critical, Marginal, and Negligible) display a mix of risk tiers. Even for negligible impacts, risk tier 2 is assigned when the likelihood is very high or high.

Table 5. Sample Risk Assessment Matrix with 4 Impact Levels, 5 Likelihood Levels, and 3 Risk Tiers

Impact	**Likelihood**				
	Very High	**High**	**Moderate**	**Low**	**Very Low**
Catastrophic	Tier 3	Tier 3	Tier 3	Tier 3	Tier 3
Critical	Tier 3	Tier 3	Tier 3	Tier 2	Tier 2
Marginal	Tier 3	Tier 2	Tier 2	Tier 1	Tier 1
Negligible	Tier 2	Tier 2	Tier 1	Tier 1	Tier 1

Table 6 highlights another example from the IEEE 1012 standard. The example in Table 5 uses 5 likelihood levels and 3 risk tiers, whereas Table 6 uses 4 likelihood levels and 4 risk tiers. This illustrates how the choices made in Steps 1-3 can change the resulting risk assessment matrix in Step 4. Table 6 also illustrates the possibility of filling in the matrix with more than one possible risk tier for a given impact and likelihood level. For example, in Table 6 critical impacts with probable likelihood map onto a risk tier of 4 or 3. This can provide organizations with additional flexibility when defining their own specific risk management programs.

Table 6. IEEE 1012's Map of 4 Impact Levels, 4 Likelihood Levels and 4 Risk Tiers (Table B.3—Graphic illustration of the assignment of integrity levels from Annex B of IEEE Standard 1012)

Impact	**Likelihood**			
	Reasonable	**Probable**	**Occasional**	**Infrequent**
Catastrophic	Tier 4	Tier 4	Tiers 4–3	Tier 3
Critical	Tier 4	Tiers 4–3	Tier 3	Tiers 2–1
Marginal	Tier 3	Tiers 3–2	Tiers 2–1	Tier 1
Negligible	Tier 2	Tiers 2–1	Tier 1	Tier 1

2.5 Step 5: Specify the risk management activities to be performed for each risk tier

The output of this section is a list of activities for managing risk in each tier. Assigning risk management activities to risk tiers represents the most complex step in the 5-step process. Table 7 shows suggested high-level themes for activities across 3 risk tiers. The NIST AI RMF and Playbook provide numerous recommendations and tasks for managing AI risk throughout an organization and across the AI lifecycle [11]. Organizations can use emergent AI standards, authoritative AI guidance from NIST and others, the organization's risk tolerance, organizational assessments of expected benefits and costs, expert judgement, domain expertise, multi-stakeholder feedback, and information relating to AI systems in similar deployment contexts to decide which risk management activities will be most appropriate for each risk tier. As shown in Table 7, a guiding principle is that higher risk tiers warrant more intensive risk management activities that take place more often, are more intensive, or require more involvement of oversight functions. To preserve risk management resources, lower risk tiers should entail less risk management activities.

Table 7. High Level Themes of Tasks/Intensity of Tasks for Risk Tiers

Risk Tier	Risk Management Tasks
3	• Intensive documentation and validation. • Extensive, frequent, and independent oversight and review. • Risk mitigation and controls across all AI RMF trustworthy characteristics.
2	• Moderate documentation and validation. • Internal oversight at a reasonable cadence. • Application of risk mitigation and controls for directly relevant AI RMF trustworthy characteristics.
1	• Nominal documentation, validation, mitigation, and oversight.

To establish AI risk tiers, organizations will typically:

- assign certain systems or applications to risk tiers based on estimated incident impact and likelihood.
- define required risk management activities for each risk tier and allocate commensurate resources.
- identify oversight roles, responsibilities, and cadences for each risk tier.

The AI RMF encourages organizations to regularly measure and document AI risk activities, as informed by interdisciplinary AI actors that are independent of the team that develops the AI technology. AI actors who both develop and evaluate AI system risks and trustworthiness have difficulty questioning their own work, and will be less effective at carrying out risk identification and management tasks. IEEE 1012 Annex C describes three aspects of independent review: technical independence, managerial independence, and financial independence and contains recommendations for the level of independence to require at different risk tiers [1]. The US Securities and Exchange Commission also provides detailed guidance for auditor independence [16].

The AI RMF core is composed of four functions: Govern, Map, Measure and Manage. Each of these high-level functions is broken down into categories and subcategories. The Govern function of the AI RMF cultivates a culture of risk management. Govern activities can operate at a high level regardless of risk tier assignment or specific AI system. The AI RMF subcategories include specified outcomes for organizations. The AI RMF is voluntary and each organization can place as many specific sub-category outcomes as they prefer into their risk tiering matrix. For example, an organization may choose to require certain actions for risk tier 3, but not for risk tier 2 or 1. Outcomes in the AI RMF, and recommended actions in the accompanying Playbook can serve as best practices for AI system risk management regardless of risk tier. Table 8 shows an example of mapping the AI RMF core functions onto a set of risk tiers. In this example, an organization prescribes a high level of governance activity regardless of risk tier, but allows the level of activity for the Map, Manage and Measure functions to vary across risk tiers.

In the Map function, context is recognized and risks related to context and context changes are identified. Many of the outcomes in the Map function align to organizational risk tolerance. Actions to meet these outcomes can help organizations assess the effectiveness of their risk tiering processes. For example, AI actors can use the contextual information captured in the Map function to determine if a risk tier 2 system is correctly assigned to that rating level, or too many systems are assigned to risk tier 2. AI RMF Map function sub-categories may be especially useful when selecting impact and likelihood levels for a system and its context of use [9]. For example Map Sub-category 5.1 states "Likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data are identified and documented." Similarly, the Map 4 category states "Risks and benefits are mapped for all components of the AI system including third-party software and data."

Table 8. A Model for Usage of the NIST AI RMF Functions Across Risk Tiers

	Govern	**Map**	**Manage**	**Measure**
Risk Tier 3	High Usage	High Usage	High Usage	High Usage
Risk Tier 2	High Usage	High Usage	Medium Usage	Medium Usage
Risk Tier 1	Medium Usage	Medium Usage	Low Usage	Low Usage

Different organizations are at different levels of maturity with respect to the AI RMF. Organizations may not be prepared to deeply engage with AI RMF details at first, but, as an organization builds up familiarity with the framework, AI actors can build up plans and activities for each category or subcategory. Operationalization of the AI RMF will require iteration and refinement over time, and risk tier guidance provides additional support for organizations using the framework to define and refine their own risk tolerance and risk tiers. Over time, organizations can establish guidelines for what is required at each risk tier for each of the 19 categories and 72 subcategories in the AI RMF. A concrete example of this level of detail is given in Table 9. Notice that the requirements are highest for Risk Tier 3 and that the requirements address not only required tasks but also required documentation and oversight.

Table 9. Sample Guidelines for Measure 2.6 Across 3 Risk Tiers

	MEASURE 2.6: The AI system is evaluated regularly for safety risks–as identified in the Map function. The AI system to be deployed is demonstrated to be safe, its residual negative risk does not exceed the risk tolerance, and it can fail safely, particularly if made to operate beyond its knowledge limits. Safety metrics reflect system reliability and robustness, real-time monitoring, and response times for AI system failures.
Risk Tier 3	• Before deployment, the AI system will be stress tested in realistic operational conditions. Tests are informed by past known failures and include efforts to test operation beyond the knowledge limits of the system. An independent team not involved in the design and implementation of the system will define a collection of stress tests, in addition to internal tests, and the independent team will review all test results. • Procedures to monitor system reliability and robustness after deployment are verified and there will be a clear process for documenting and responding to AI system failure. • Processes for any impacted individual to report negative impacts are verified and records of these reports and responses to them will be maintained. • Readiness of redundant systems is verified. • Substantial model documentation and testing of incident response plans.
Risk Tier 2	• Before deployment, the AI system will be tested in realistic operational conditions. It is recommended that an independent team not involved in the design and implementation of the system be involved in testing. • Processes for documenting and responding to AI system failure are verified. • Processes for any impacted individual to report negative impacts are verified and records of these reports and responses to them will be maintained. • Documentation of model and of incident response plans.
Risk Tier 1	• Processes for any impacted individual to report negative impacts are verified and records of these reports and responses to them will be maintained. • Minimal documentation of model and of incident response plans.

As an organization's AI risk management capabilities mature, Table 9 could be expanded to include requirements for change management plans, accelerated shutdown plans, manual restrictions on system behavior, and other risk mitigation and controls. The AI RMF Playbook [11] also contains suggested actions, references, and documentation guidance to achieve outcomes for the four AI RMF functions. Organizations can consider including specific items from the AI RMF Playbook for any risk tier.

3. Using Risk Tiers

Section 2 presented a practical process for defining a set of risk tiers at the organizational level, discussed basic ideas related to assigning resources, requirements, applications, or systems to risk tiers, and addressed mapping the AI RMF sub-category outcomes onto risk tiers. Section 3 continues the discussion on risk tier assignment and covers a few additional aspects of how organizations may use risk tiers across portfolios of AI systems.

3.1 Using Organizational Risk Tiers For Different AI Systems

Once risk tiering is in place, an organization can then assess their full portfolio of AI systems, existing and proposed, within this context. Organizations with identified systems that fall into the lowest risk tier may decide to not invest substantial risk management resources into those systems and instead focus risk management activities and allocate resources to systems that fall into higher risk tiers. An organization may discover that an existing system falls into a high risk tier and then decide to work on improving risk management capabilities. An organization may evaluate a proposed system and delay development until it can improve its risk management posture. Organizations may revise risk tier assignment of resources, requirements, applications, or systems to align the risk management activities with available resources, gain efficiency, or improve adherance to organizational risk tolerances. Implicit in assigning all AI systems to risk tiers is taking an inventory of all organizational AI systems. AI inventories are an important risk management and governance tool.

3.2 Risk Tiers and AI Inventories

AI RMF Govern Sub-category 1.6 advises organizations to store relevant information about AI systems in an organized repository, and augment inventoried information with model risk tiers. Hence, risk tiers should inform the design of model inventories and associated policies and procedures. Generally, higher risk systems require more detailed documentation, human review and oversight, incident response plans, and associated metadata to facilitate risk management, and lower risk systems require less resources. These materials can be stored, along with other system artifacts in an inventory. When a system's risk tier is also stored in a structured inventory, risk can be tracked at the system, division, and enterprise levels. By coupling risk tiers and inventories as part of broader AI governance and risk management efforts, risk can be measured for specific systems, or aggregated to various levels within the organization. Aggregating risk across all AI systems enables organizations to verify that AI risk is aligned with board and senior management risk tolerance statements.

3.3 Productivity Tools and Vendor Products

It is advisable that all AI systems–whether open source, built in house, or provided by a vendor–that are used for decision-making, enterprise-level generative tasks, or any other material matter are included in risk tiers and formal risk management activities. However, organizations with established model governance programs often increase efficiency in risk management by designating certain tools and vendor products as outside the direct purview of AI risk management activities. Some tools that implicate models or AI, e.g., dashboards, spreadsheets, or chatbots embedded in standard productivity software, may be inefficient to oversee from a risk management perspective and can be excluded from risk tiers–particularly when vendors are able to provide ample evidence of risk management for their implementation and maintenance of the tool. If vendors cannot provide evidence of risk management, it is likely appropriate for their tools to be assigned to a risk tier. Under some AI governance programs, risks arising from productivity tools and certain low-risk vendor tools are managed with thorough acquisition and procurement controls, legal review of contracts, documented individual ownership of vendor relationships, end-user or responsible use policies, and minimal documentation included in model inventories. As certain applications of productivity tools and vendor software can be high-risk, inclusion in inventories provides oversight functions with awareness of the software. Oversight or governance functions may elect to re-certify exclusion from risk tiers by auditing tools, products, and applications at reasonable cadences.

3.4 Updating Risk Tiers Based on Post-Deployment Data and Feedback

Once AI systems are deployed, risk tier settings can be updated based on monitoring data for negative impacts. Related data can be provided by external stakeholders, who can communicate detailed insights about impacts to AI actors to improve organizational assessments and risk tier processes. Post-deployment data may reveal insights about real performance that were missed in organizational risk estimates and impact assessments. The collection of post-deployment feedback and error reports is worthwhile, even for systems estimated to be at the lowest risk tier, to enable organizations to adjust or confirm their risk management activities in relation to on-the-ground realities. These activities relate to subcategory Map 5.2 of the AI RMF ("Practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts are in place and documented") [9]. Multi-stakeholder feedback processes from internal or external groups may also help improve risk tiers. Groups of interest may include internal business or technology functions that undergo validation and oversight in accordance with risk tiers, or external impacted communities that can provide crucial insights into frequency and impacts of any adverse events. Like AI systems themselves, AI risk management can benefit from the perspective of many different AI actors.

3.5 Different Risk Perspectives for Different AI Actors

Measuring risk at an earlier stage in the AI lifecycle may yield different results than measuring risk at a later stage. Some risks may be latent at a given point in time and may increase as AI systems adapt and evolve. Different AI actors across the AI lifecycle can have different risk perspectives. For example, an AI developer who makes AI software available, such as pre-trained models, can have a different risk perspective than an AI actor who is responsible for deploying that pre-trained model in a specific use case. Such deployers may not recognize that their particular uses could entail risks which differ from those perceived by the initial developer. All involved AI actors share responsibilities for designing, developing, and deploying a trustworthy AI system that is fit for purpose thoughtout its lifecycle [9].

Using risk tiers is an iterative processes that is connected to several aspects of broader AI governance and enterprise risk management. For the best risk management outcomes, risk tiers can be customized to organizational needs, refined over time based on monitoring of deployed AI systems, and enhanced with feedback from a wide range of stakeholders. For many organizations, instating risk tiers for their AI portfolio is a novel exercise. Other organizations have used structured tiers for many years in risk management of AI and other autonomous systems. Section 4 below surveys risk tiers to provide examples and information about existing approaches.

4. Survey

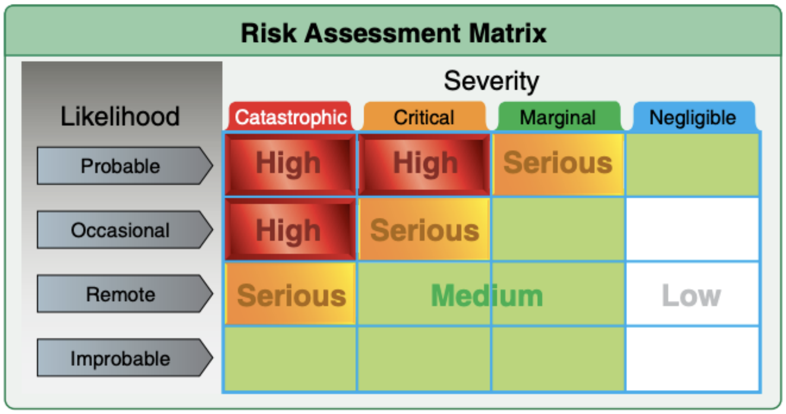
Risk management practices from a variety of real-world sources can provide practical examples for organizations seeking to set up their own risk tiering processes. Section 4 surveys examples from US federal government agencies, international standards bodies, private industry, and the EU AI Act. These examples illustrate how risk tiering can take different forms.

4.1 US Federal Aviation Administration

The FAA follows a risk tiering process much like the one presented in this document, including a range of failure impact levels from high to low and a range of failure likelihood levels from high to low. Risk management activities to be performed are specified at each risk tier. The process is documented in the FAA Risk Management Handbook [2]. The FAA defines risk severity or impact levels specific to the aviation context. Harms such as airframe loss or damage are specified, as shown in Table 2. This risk tier matrix uses 4 terms to describe risk likelihood: probable, occasional, remote or improbable. It also presents risk as a composite of the likelihood (probability) and the severity (impacts) of a particular outcome where both likelihood and severity can vary in magnitude.

The FAA 2-D Risk Assessment Matrix displayed in Figure 1 and maps four risk levels: red/high, yellow/serious, green/medium, and white/low. A risk that falls into the high (red) or serious (yellow) categories implies “no-go” unless the pilot finds a means to reduce the risk, enough so that the next iteration of risk assessment indicates a medium or low risk. While the medium (green) risk level implies a “go” decision for a planned or ongoing flight, risk should still be mitigated as applicable.

The FAA Risk Management Handbook describes common errors when using a matrix



Risk Assessment Matrix

	Severity			
Likelihood	Catastrophic	Critical	Marginal	Negligible
Probable	High	High	Serious	
Occasional	High	Serious		
Remote	Serious	Medium		Low
Improbable				

Fig. 1. Figure 4-1 from Chapter 4, "Assessing Risk" in the FAA Risk Management Handbook [2]

to assign risk levels. For example, a pilot may be unsure of the likelihood or severity for a particular risk or find it difficult to choose between adjacent parameters. In these cases, the Handbook instructs pilots to apply the more conservative parameter(s), resulting in a higher risk level. The Handbook also describes a practice where pilots often use inputs that skew the risk toward a lower level, due to their desire to complete a flight. Pilots are encouraged to verify and reevaluate their risk assessment as conditions change (e.g., based on current information and conditions before flight).

Another informative aspect of transportation risk management includes a database of incidents with root cause analysis. The National Transportation Safety Board (NTSB) maintains a searchable aviation accident database containing civil aviation accidents and selected incidents, from 1962 to present within the United States, its territories and possessions, and in international waters [17]. A number of similar AI-focused databases exist including the AI Incident Database [12], the AIAAIC [13], and the AI Vulnerability Database [14]. Consulting these and similar resources can help assign likelihood, impact, and risk levels, and assist AI actors in avoiding similar future incidents.

4.2 US Food and Drug Administration

The US FDA offers another relevant example of risk tiering. Risks in this case are not defined by severity of impacts and the probability of those impacts. Instead, tiers are defined related to the required risk controls (general controls, special controls, more than special controls). Medical devices are assigned to tiers based on the risk the device poses to patients and users. The tiers prescribe the level of control necessary to assure the safety and effectiveness of the device [15]. The FDA's three classes (or risk tiers) are:

- **Class I:** Devices are subject to a comprehensive set of regulatory authorities called general controls that are applicable to all classes of devices

- **Class II:** Devices for which general controls, by themselves, are insufficient to provide reasonable assurance of the safety and effectiveness of the device, and for which there is sufficient information to establish special controls to provide such assurance.
- **Class III:** Devices for which general controls, by themselves, are insufficient and for which there is insufficient information to establish special controls to provide reasonable assurance of the safety and effectiveness of the device. Class III devices typically require premarket approval.

The FDA device classification depends on the intended use and specialized indications for use of the device. For example, a scalpel's intended use is to cut tissue. A subset of intended use arises when a more specialized indication is added in the device's labeling such as, "for making incisions in the cornea" [15]. Class I includes devices with the lowest risk and Class III includes those with the greatest risk. The class to which a device is assigned determines, among other things, the type of premarket submission/application required for FDA clearance to market. Non-exempt Class I or II devices require premarket notification (a 501(k) submission) prior to marketing the device [18]. Class III devices typically require more extensive premarket approval [15].

FDA has further classified over 1,700 distinct types of devices into 16 medical specialty "panels" such as Cardiovascular devices and Ear, Nose, and Throat devices. Medical device manufacturers can search device classification panels to determine their device class and any exemptions. If there are no exemptions, manufacturers can also request a formal device determination or classification from the FDA for a fee (with reduced fees for small businesses) [19]. The International Medical Device Forum (IMDF) specifically addresses software as a medical device (SaMD) and offers a framework for risk categorization in that context [20]. Some SaMD systems rely on models or AI systems. The IMDF lays out 4 categories–I, II, III, IV–based on the levels of impact on the patient or public health. Any manufacturer's SaMD definition statement claiming use across multiple healthcare situations or conditions requires categorization at the highest risk level. Table 10 summarizes SaMD risk tiers.

Table 10. SaMD Categories

State of healthcare situation or condition	**Significance of information provided by SaMD to healthcare decision**		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	Category IV	Category III	Category II
Serious	Category III	Category II	Category I
Non-Serious	Category II	Category I	Category I

4.3 International Standards

The IEEE, a technical professional association, maintains established processes and standards to verify and validate critical systems such as United States Department of Defense

(DoD) weapons systems, nuclear weapons systems, power control systems, space vehicles, and medical devices. For these verification and validation processes, the IEEE applies a process similar to the 5 steps proposed in this document to classify risk of software and hardware systems, including AI systems [1], [21]. IEEE Standard 1012 establishes a range of failure impact levels from high to low, a range of failure likelihood levels from high to low, and number of integrity levels (similar to risk tiers) from high to low. Impacts are mapped to likelihood and integrity levels using a 2-D matrix. Activities at each integrity level correspond to verification and validation outputs such as documentation of requirements, design, testing, and maintenance activities. International Organization for Standardization (ISO) also puts forward numerous standards related to AI risk management. ISO 31000 addresses general risk management and ISO 23894 applies directly to AI risk management [6, 22]. ISO Guide 73-2009 defines levels of risk and risk matrices, but does not provide specific examples of risk tiers [4].

4.4 Private Industry Practices

Many large US consumer finance institutions implement model risk management practices under the interagency supervisory guidance of the Federal Reserve Bank and other financial regulators [23], [24]. While such guidance rarely specifies specific risk tiers, it does reference:

- The notion of materiality as the product of impact and likelihood of adverse events relating to systems.
- Ongoing risk assessments at system and organization levels.
- Measuring risks based on system type, objectives, complexity, uncertainty, materiality, interrelationships, data, capabilities, and limitations.
- Risk measurements informing the types, frequency, and extent of validation activities and allocated risk management resources.

More recently, financial institution risk guidance has focused on AI system explainability as a criterion for risk measurement. The 2021 OCC Examiner Handbook specifies that explainability should inform risk assessments [24], and the US Consumer Financial Protection Bureau (CFPB) has highlighted risks associated with the use of complex, and difficult to explain, AI systems in the provisioning of consumer credit [25], [26].

Given their regulatory and supervisory requirements, most large US banks construct risk tiers for AI systems used in various organizational decision support tasks. The highest risk tiers tend to apply to systems used for credit underwriting, loss forecasting, other systems that directly impact revenue and reserves, or systems or applications that undergo serious regulatory scrutiny. Middle risk tiers often apply to systems used for fraud detection or large marketing campaigns, and the lowest tiers likely include models used for administrative, back-office, self-service, or additional internal or low-materiality applications.

Certain vendor tools and productiticy tools with embedded analytical or AI capabilities may be excluded from risk tiers and formal model risk management, with risk managed via procurement processes, vendor relationship owners, and minimal documentation.

Banks often assign systems to one of 3-5 risk tiers, and they may use expert judgment, decision trees, scorecards, business rules, internal data, or other approaches to assign models to risk tiers [7]. Different risk tiers often encompass different documentation, review and validation requirements. Systems in the highest risk tiers tend to undergo in-depth validation and human review, with copious documentation requirements for development and validation teams. Validation for high-risk tier systems may require months of testing and remediation, with hundreds of pages of system documentation generated in the process, and multiple layers of management or internal audit review. Conversely, systems in the lowest tiers receive substantially less human review and validation treatment, while also requiring less thorough documentation.

Another example of risk tiering in private industry is Anthropic's Responsible Scaling Policy [27]. It proposes a series of AI Safety Levels: ASL-1 through ASL-5+. ASL-1 refers to systems which pose no meaningful catastrophic risk, such as older language models, or an AI system that only plays chess. ASL-2 refers to systems that may output harmful information–for example, the ability to give instructions on how to build bioweapons–but not any more information than might be available from an online search engine. Large language models in 2023 are generally considered to be ASL-2. ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g., search engines or textbooks) or systems that are perceived to display low-level autonomous capabilities. ASL-4 and higher (ASL-5+) is not yet defined.

4.5 EU AI Act

The risk tiers in the European Union's draft AI Act include unacceptable risk, high risk, and limited or minimal risk [28]. Unlike the approach described in this document, the EU AI Act risk tiering process focuses more on identifying high risk technologies or use cases and assigning such systems directly to a tier. The AI Act prohibits certain AI applications, including behavior modification resulting in physical or psychological harm, subliminal techniques, predatory techniques for persons due to age, physical or mental disability, social credit scoring, and real-time biometric identification (with exceptions for law enforcement or anti-terrorism activities) [28] (Article 5). The EU AI Act high-risk tier applies to systems used as a safety component of another product, biometric identification and categorization of natural persons, and systems used in management and operation of critical infrastructure, education and vocational training, employment, worker management and access to self-employment, access to and enjoyment of essential private services and public services and benefits, law enforcement, migration, asylum and border control management, and the administration of justice and democratic processes [28] (Article 6 and Annexes). Systems in the high risk tier must meet formidable validation, documentation, and audit standards to ensure conformity with the Act. Risk management requirements are lower for systems assigned to lower tiers.

5. Conclusion

Organizations using the NIST AI Risk Management Framework (AI RMF) are encouraged to define a reasonable organizational risk tolerance for AI and apply it to their specific use cases. Risk tiering is a common method to do so. To help organizations select from the many available risk tiering options, this document specifies a practical process based on the severity of possible impacts and the likelihood of those impacts occurring. This document also provides a landscape survey of risk tiering and other risk management practices from government agencies, international standards bodies, and private industry. Defining reasonable risk tolerance and risk tiers is iterative and requires assessment and refinement over time. Organizations may use the processes or the survey materials herein to inform their own risk tiers and increase alignment with the voluntary NIST AI RMF.

References

[1] Institute of Electrical and Electronics Engineers (IEEE) (2016) IEEE Standard for System, Software, and Hardware Verification and Validation, IEEE 1012-2016, 2016.

[2] US Department of Transportation, Federal Aviation Administration (FAA) (2022) Risk Management Handbook (FAA-H-8083-2A). Available at https://www.faa.gov/regulationspolicies/handbooksmanuals/risk-management-handbook-faa-h-8083-2a.

[3] National Institute of Standards and Technology (NIST) (2012) Guide for Conducting Risk Assessments (NIST Special Publication 800-30). Available at https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf.

[4] International Organization for Standardization (ISO) (2009) Guide 73: Risk Management - Vocabulary, ISO 73:2009.

[5] National Institute of Standards and Technology (NIST) (2023) NIST AI RMF Glossary. Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary.

[6] International Organization for Standardization (ISO) (2018) Risk Management - Guidelines, ISO 31000:2018.

[7] Nick Kiritz, Miles Ravitz, Mark Levonian (2019) Model Risk Tiering: An exploration of industry practices and principles. Available at https://www.risk.net/journal-of-risk-model-validation/6710566/model-risk-tiering-an-exploration-of-industry-practices-and-principles.

[8] International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) (2022) Trustworthiness Vocabulary, ISO/IEC TS 5723:2022.

[9] National Institute of Standards and Technology (NIST) (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0) (US Department of Commerce, Washington, D.C.), Artificial Intelligence Publications (AI PUBS) 100-1, January 2023. https://doi.org/10.6028/NIST.AI.100-1

[10] Office of the Comptroller of the Currency (OCC), and the Federal Deposit Insurance Corporation (FDIC) (2020) Sound Practices to Strengthen Operational Resilience. Available at https://www.occ.gov/news-issuances/news-releases/2020/nr-occ-2020-144a.pdf.

[11] National Institute of Standards and Technology (NIST) (2023) NIST AI RMF Playbook. Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook.

[12] AI Incident Database. Available at https://incidentdatabase.ai/.

[13] AIAAIC (AI, Algorithmic, and Automation Incidents and Controversies). Available at https://www.aiaaic.org/.

[14] AI Vulnerability Database. Available at https://avidml.org/database/.

[15] US Food and Drug Administration (FDA) (2022) Classify Your Medical Device. Available at https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device.

[16] US Securities and Exchange Commission (SEC) (2023) Guidance for Auditor Independence. Available at https://www.sec.gov/page/oca-independence-guidance.

[17] National Transportation Safety Board (NTSB) (2023) Aviation Investigation Search. Available at https://www.ntsb.gov/Pages/AviationQueryV2.aspx.

[18] US Food and Drug Administration (FDA) (2014) The 510(k) Program: Evaluating Substantial Equivalence in Premarket Notifications [510(k)] Guidance for Industry and Food and Drug Administration Staff. Available at https://www.fda.gov/media/82395/download.

[19] US Food and Drug Administration (FDA) (2023) Medical Device User Fee Amendments (MDUFA). Available at https://www.fda.gov/industry/fda-user-fee-programs/medical-device-user-fee-amendments-mdufa.

[20] IMDRF Software as a Medical Device (SaMD) Working Group (2014) Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations. Available at https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf.

[21] Jeanna Matthews, Bruce Hedin, Marc Canellas (2022) Trustworthy Evidence for Trustworthy Technology: An Overview of Evidence for Assessing the Trustworthiness of Autonomous and Intelligent Systems. Available at https://ieeeusa.org/assets/public-policy/committees/aipc/IEEE_Trustworthy-Evidence-for-Trustworthy-Technology_Sept22.pdf.

[22] International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) (2023) Artificial Intelligence, Guidance on Risk Management, ISO/IEC 23894:2023(en), 2023.

[23] Federal Reserve Bank (2011) Interagency Guidance on Model Risk Management. Available at https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

[24] Office of the Comptroller of the Currency (2021) Model Risk Management: Comptroller's Handbook. Available at https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.

[25] Consumer Financial Protection Bureau (2022) CFPB Acts to Protect the Public from Black-Box Credit Models Using Complex Algorithms. Available at https://www.consumerfinance.gov/about-us/newsroom/cfpb-acts-to-protect-the-public-from-black-box-credit-models-using-complex-algorithms/.

[26] Consumer Financial Protection Bureau (2023) CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence. Available at https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/.

[27] Anthropic (2023) Anthropic's Responsible Scaling Policy. Available at https://www.anthropic.com/index/anthropics-responsible-scaling-policy.

[28] European Union AI Act. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.